

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**



THOMSON  
DELPHION

RESEARCH

Log Out | Work Files | Saved Searches | My Account | Products

PRODUCTS

INSIDE DELPHION

Search: QuickNumber | Boolean | Advanced | Delwait | Help

## The Delphion Integrated View

Buy Now: ☒ PDF | [More choices...](#)

Tools: Add to Work File:

View: INPADOC | Jump to:

Title: JP2003099384A2: LOAD-SHARING SYSTEM, HOST COMPUTER FOR THE LOAD-SHARING SYSTEM, AND LOAD-SHARING PROGRAM

Country: JP Japan  
Kind: A2 Document Laid open to Public inspection  
Inventor: KATSURAJIMA KO;  
Assignee: NEC CORP  
News, Profiles, Stocks and More about this company

Published / Filed: 2003-04-04 / 2001-09-20

Application Number: JP2001000286952

IPC Code: G06F 13/14; G06F 3/06; G06F 12/08;

Priority Number: 2001-09- JP2001000286952

Abstract: PROBLEM TO BE SOLVED: To provide a multi-path load-sharing system in a host computer, which accesses a data storage device with unshared cache method.

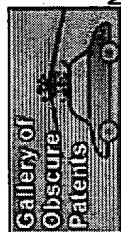
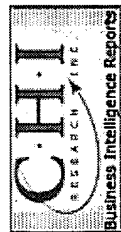
SOLUTION: In the load-sharing system, connecting a host computer 100 and a data storage device 200 with the unshared cache method via a network 300, a host computer 100 is constituted with a load-sharing control part 130 which groups plural physical paths with the data storage device 200 for each of caches which are to be used by a host control part, for executing prefetch within the data storage device 200 and manage them as a path group, map a sequential readout request into a single path group and properly distributing input output requests into the plural



physical paths.  
COPYRIGHT: (C)2003,JPO

Family: None

Other Abstract Info: DERABS G2003-325008



[Nominate this for the Gallery...](#)



© 1997-2004 Thomson    Research Subscriptions | Privacy Policy | Terms & Conditions | Site Map | Contact Us | Help

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2003-99384

(P2003-99384A)

(43) 公開日 平成15年4月4日 (2003.4.4)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テマコード* (参考)
G 0 6 F 13/14	3 1 0	G 0 6 F 13/14	3 1 0 H 5 B 0 0 5
3/06	3 0 2	3/06	3 0 2 A 5 B 0 1 4
12/08	5 1 1	12/08	5 1 1 Z 5 B 0 6 5
	5 1 5		5 1 5 Z
	5 1 9		5 1 9 Z

審査請求 未請求 請求項の数22 O L (全 17 頁) 最終頁に続く

(21) 出願番号 特願2001-286952(P2001-286952)

(22) 出願日 平成13年9月20日 (2001.9.20)

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 桂島 航

東京都港区芝五丁目7番1号 日本電気株

式会社内

(74) 代理人 100093595

弁理士 松本 正夫

Fターム(参考) 5B005 JJ13 MM12 NN12

5B014 HA13

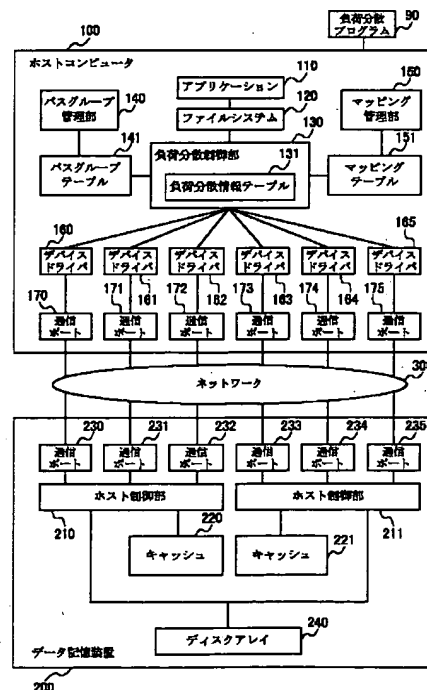
5B065 CC08 CE13 ZA08

(54) 【発明の名称】 負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラム

(57) 【要約】 (修正有)

【課題】 非共有キャッシュ方式のデータ記憶装置にアクセスするホストコンピュータにおいてマルチパス負荷分散システムを提供する。

【解決手段】 ホストコンピュータ100と、非共有キャッシュ方式のデータ記憶装置200をネットワーク300を介して接続する負荷分散システムにおいて、ホストコンピュータ100は、データ記憶装置200との間における複数の物理的パスを、データ記憶装置200内においてプリフェッチを実行するホスト制御部が使用するキャッシュ毎にグループ化してパスグループとして管理し、シーケンシャルな読み出し要求を単一のパスグループにマッピングし、入出力要求を複数本の物理的パスに適切に分散する負荷分散制御部130を備える。



## 【特許請求の範囲】

【請求項1】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、  
入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、  
前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする負荷分散システム。

【請求項2】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、  
前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理し、前記バスグループの情報を示すバスグループ情報を生成し、  
入出力要求のアドレス情報と前記バスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、  
前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする負荷分散システム。

【請求項3】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、  
前記データ記憶装置と前記ホストコンピュータとの間の通信を中継し、前記データ記憶装置のキャッシングを実行するホスト制御装置を備え、  
入出力要求のアドレス情報と、前記ホストコンピュータと前記ホスト制御装置との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、  
前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする負荷分散システム。

【請求項4】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、  
前記データ記憶装置と前記ホストコンピュータとの間の通信を中継し、前記データ記憶装置のキャッシングを実行するホスト制御装置を備え、  
前記ホストコンピュータと前記ホスト制御装置との間における複数の物理的バスを、前記ホスト制御装置内のキャッシュ毎にグループ化してバスグループとして管理し、前記バスグループの情報を示すバスグループ情報を

生成し、

入出力要求のアドレス情報と前記バスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、  
前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする負荷分散システム。

【請求項5】 前記マッピング情報の生成において、前記入出力要求のアドレス情報として、当該入出力要求の論理ボリューム番号を用いることを特徴とする請求項1から請求項4のいずれか1つに記載の負荷分散システム。

【請求項6】 前記入出力要求の論理アドレス番号を適当な幅毎に区画化し、  
前記マッピング情報の生成において、前記入出力要求のアドレス情報として、当該入出力要求の論理アドレス番号が該当する前記区画を用いることを特徴とする請求項1から請求項4のいずれか1つに記載の負荷分散システム。

【請求項7】 前記ネットワークを介して、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする請求項6記載の負荷分散システム。

【請求項8】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、  
前記ホストコンピュータが、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理するバスグループ情報を記録するバスグループテーブルと、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングした、入出力要求のアドレス情報と前記バスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルを備え、  
前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする負荷分散システム。

【請求項9】 非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、  
入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理手段と、  
前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御手段を備えることを特徴とするホストコンピュータ。

【請求項10】 非共有キャッシュ方式のデータ記憶装

置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理し、前記バスグループの情報を示すバスグループ情報を生成するバスグループ管理手段と、入出力要求のアドレス情報と前記バスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理手段と、前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御手段を備えることを特徴とするホストコンピュータ。

【請求項11】 前記バスグループ管理手段は、前記データ記憶装置から、当該データ記憶装置が記憶するデータの構成情報を取得する手段を備えることを特徴とする請求項10に記載のホストコンピュータ。

【請求項12】 前記バスグループ管理手段は、指定された外部の情報機器から、前記データ記憶装置が記憶するデータの構成情報を取得する手段を備えることを特徴とする請求項10に記載のホストコンピュータ。

【請求項13】 前記マッピング管理手段は、前記入出力要求のアドレス情報として、当該入出力要求の論理ボリューム番号を用いてマッピングを行なうことを特徴とする請求項9から請求項12のいずれか1つに記載のホストコンピュータ。

【請求項14】 前記マッピング管理手段は、前記入出力要求の論理アドレス番号を適当な幅毎に区画化し、前記入出力要求のアドレス情報として、当該入出力要求の論理アドレス番号が該当する前記区画を用いてマッピングを行なうことを特徴とする請求項9から請求項12のいずれか1つに記載のホストコンピュータ。

【請求項15】 前記マッピング管理手段は、前記データ記憶装置から、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする請求項14記載のホストコンピュータ。

【請求項16】 前記マッピング管理手段は、指定された外部の情報機器から、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする請求項14記載のホストコンピュータ。

【請求項17】 前記外部の情報機器を、ネームサーバとすることを特徴とする請求項12又は請求項16に記載のホストコンピュータ。

【請求項18】 前記マッピング管理手段は、

定期的に各前記バスの負荷情報を参照して前記マッピング情報を更新することを特徴とする請求項9から請求項17のいずれか1つに記載のホストコンピュータ。

【請求項19】 非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、

前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理するバスグループ情報を記録するバスグループテーブルと、

シーケンシャルな読み出し要求を単一の前記バスグループにマッピングした、入出力要求のアドレス情報と前記バスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルと、

前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御手段を備えることを特徴とするホストコンピュータ。

【請求項20】 非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、

入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理機能と、

前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御機能を備えることを特徴とする負荷分散プログラム。

【請求項21】 非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、

前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理し、前記バスグループの情報を示すバスグループ情報を生成するバスグループ管理機能と、

入出力要求のアドレス情報と前記バスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理機能と、

前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御機能を備えることを特徴とする負荷分散プログラム。

【請求項22】 非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータ

を制御することにより、負荷分散を実行する負荷分散プログラムにおいて、前記ホストコンピュータが、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理するバスグループ情報を記録するバスグループテーブルと、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングした、入出力要求のアドレス情報と前記バスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルを備え、前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御機能を備えることを特徴とする負荷分散プログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明はホストコンピュータとデータ記憶装置の間に複数本の物理的アクセスバスが存在する場合のホストコンピュータのマルチバス負荷分散に関し、特に、データ記憶装置が非共有キャッシュ方式の場合の負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラムに関する。

【0002】

【従来の技術】高可用性を求められるシステムの場合、ホストコンピュータとデータ記憶装置のアクセスのために、複数本の物理的バスが設けられる。その理由は、或るバスで障害が発生した際に、異なるバスをその代替バスとして用いるためである。

【0003】これらの複数本の物理的バスを効率的に利用するために、ホストコンピュータ側でのマルチバス負荷分散が広く行なわれている。データ記憶装置に対するアクセスの負荷を複数本の物理的バスに適切に分散することで、ホストコンピュータとデータ記憶装置間のスループットを向上させるのが狙いである。

【0004】マルチバス負荷分散の制御を行なう手段は、ファイルシステムレイヤとデバイスドライバレイヤの間に位置しており、ホストコンピュータのファイルシステムからデータ記憶装置に対するI/O要求(入出力要求)を受け取ると、負荷分散アルゴリズムに基づき使用するバスを決定し、適切なデバイスドライバへI/O要求を渡す仕組みになっている。

【0005】次に、図面を参照して従来の負荷分散システムの動作を説明する。図12は、従来の負荷分散システムの構成を示すブロック図である。

【0006】ホストコンピュータ100e内のファイルシステム120は、アプリケーション110からI/O要求を受け取ると、それらをブロック単位のI/O要求に変換し、負荷分散制御部130eに送付する。負荷分散制御部130eは、ファイルシステム120からI/O要求を受け取ると、負荷分散アルゴリズムに基づき使

用するバスを決定する。最後に、デバイスドライバ160~165の中で適切なものに対し、I/O要求を送付する。

【0007】負荷分散のアルゴリズムとしては、発生するI/O要求を、複数本のバスに対しラウンドロビンにて送付する方法、ベンディングされているI/O要求の数が最も少ないバスに送付する方法、ベンディングされているI/O要求のブロック数の総和が最も少ないバスに送付する方法等が使われている(ここでベンディングされているI/O要求とは、確認応答が返ってきていないI/O要求のことを指している)。

【0008】また、従来技術の一例として、特開平07-334449号公報に開示された技術が挙げられる。この方式は、物理的バスを識別する入出力経路識別部と、各物理的バスの負荷を測定する負荷測定部と、負荷情報を蓄積する負荷統計情報蓄積部と、蓄積された負荷情報に基づきバスを決定する入出力経路決定部を備える。

【0009】入出力経路決定部は、周辺装置に対する入出力要求時に指定される閾値と、前記負荷統計情報蓄積部によって蓄積された前記統計情報によって示される入出力制御装置の負荷状況とを比較した結果に基づいて、入出力経路の変更が必要か否かを判別する情報提供部と、前記情報提供部によって、負荷状況が閾値を越えて入出力経路の変更が必要と判別された場合に、複数の入出力経路から低負荷の経路を選択する低負荷経路選択部とを具備し、前記低負荷経路選択部によって選択された入出力経路を経由して入出力要求が行なわれることを特徴とする。

【0010】

【発明が解決しようとする課題】しかし、上述した従来の負荷分散システムでは、非共有キャッシュ方式のデータ記憶装置に対してマルチバス負荷分散を用いると、データ記憶装置のプリフェッチ性能が低下するという問題点があった。以下、この問題点について詳細に説明する。

【0011】高機能なデータ記憶装置は、ホストコンピュータからのI/O要求をRAID (Redundant Array of Inexpensive Disks) によるディスクアレイで実行する前に、プロセッサ等を用いて処理を行なうことでいくつかの機能を実現している。以下、この処理を行なう部分をホスト制御部と呼ぶことにする。

【0012】ホスト制御部の機能としては、ホストコンピュータ側に提示するボリュームを柔軟に構成することを可能にする仮想ボリューム機能、スループットやレスポンス向上のためにRAM (Random Access Memory) 等を使って行なわれるキャッシング機能、等が挙げられる。

【0013】キャッシングは、ホストコンピュータに対して磁気ディスクへのアクセス時間を隠蔽することでス



ループットやレスポンスタイムを向上させる機能である。すなわち、リード要求はキャッシュヒットすれば磁気ディスクにアクセスする必要がなくなるし、ライト要求はとりあえずキャッシングしてその時点でホストコンピュータへ確認応答を返すようにすれば、ホストコンピュータ側からは磁気ディスクへのアクセスはまるで起こっていないかの様に見える。

【0014】ホスト制御部は、並列に複数個備えられることが多い。その理由は2つある。

【0015】1つ目は高可用性である。すなわち、或るホスト制御部で障害が発生した際に、異なるホスト制御部をその代替として用いるためである。

【0016】2つ目は、ボトルネック防止である。すなわち、ホスト制御部の処理能力がシステムの他の能力と比べて低くならないように、同様の処理が行えるホスト制御部を並列に複数配置して処理能力の向上を図るわけである。

【0017】このように複数のホスト制御部を持つデータ記憶装置の場合、キャッシュの構成方式によって共有キャッシュ方式と非共有キャッシュ方式の2つに分類できる。図13は、従来の共有キャッシュ方式のデータ記憶装置200fと、非共有キャッシュ方式のデータ記憶装置200gの構成を示すブロック図である。

【0018】全てのホスト制御部から単一のキャッシュ、単一のキャッシュ制御情報を参照する方式が共有キャッシュ方式であり、いくつかのホスト制御部毎に異なるキャッシュ、異なるキャッシュ制御情報を参照する方式が非共有キャッシュ方式である。

【0019】ここでキャッシュ制御情報とは、キャッシュしたデータのアドレス情報やキャッシングの順序関係を整理して記した情報のことであり、どのデータをキャッシングするか、又はどのデータを掃き出すかを決定するための情報として使われる。

【0020】共有キャッシュ方式では、同じアドレス領域のデータを二重にキャッシュすることがないためにキャッシュ容量を有効に使うことができ、またキャッシュは1つだけ用意すればよいから容量を比較的大きくすることができるという利点がある。結果として、キャッシュヒット率が非共有キャッシュ方式に比べて高くなる。欠点は、複数のプロセッサによる競合が発生するため遅延が若干発生することと、複雑な制御を行なうためコストが高くなりがちなことである。

【0021】一方非共有キャッシュ方式は、複数のプロセッサによる競合が発生しにくい、構造が簡単になりコストを低くできる等の利点がある反面、キャッシュの一貫性の維持という問題がある。

【0022】この問題の対応策としては、キャッシュ間でダーティデータのみコピーを行なう方法が一般的である。ここでダーティデータとは、後に磁気ディスクにライトされる目的でキャッシングされているデータのこと

を指している。どこかのキャッシュにデータが書き込まれダーティデータとなった場合、他のキャッシュにコピーを行なうのである。この方法によって、キャッシュの一貫性を保つことができる。

【0023】また、非共有キャッシュ方式では、キャッシュ制御情報がキャッシュ毎に別々に扱われているために、プリフェッチ等の機能がキャッシュを共有するホスト制御部群を単位として働く。

【0024】プリフェッチ機能とは、ホストコンピュータからのI/O要求の中からシーケンシャルなリードI/O（読み出し要求）を選び分け、それらに関しては先回りして磁気ディスクからデータを読み出してキャッシュに蓄えておくという機能である。具体的には、近接するアドレスのリードI/Oが或る程度連続してデータ記憶装置に到着した場合、キャッシュ制御情報からそれらのI/OがシーケンシャルなリードI/Oであると判断し、以降のアドレスのデータを適当な分だけ読み出してキャッシュに蓄える。

【0025】例えば、ホストコンピュータがシーケンシャルなリードI/Oをデータ記憶装置に送った場合、ホスト制御部は初めの3、4個程度の近接するアドレスのリードI/Oを受け取った時点でそれ以降のアドレスに対するリードI/Oを磁気ディスクに送付する。順次ホストコンピュータから到着するリードI/Oは、キャッシュヒットしレスポンスタイムが大幅に削減されるという仕組みである。

【0026】本発明が解決する課題は、非共有キャッシュ方式のデータ記憶装置に対してマルチバス負荷分散を用いると、データ記憶装置のプリフェッチ性能が著しく低下するという問題である。

【0027】例えば、マルチバス負荷分散の負荷分散アルゴリズムにラウンドロビンを用いている場合、ホストコンピュータで発生したシーケンシャルなリードI/Oは各バスに対して順番に振り分けられるから、複数のキャッシュ制御情報にそれぞれ断続的なアドレスへのリードI/Oが記録されることになる。この場合、性能低下を招く要因が2点存在する。

【0028】1点目は、キャッシュ制御情報に断続的なアドレスのリードI/Oが記録されにくいために、シーケンシャルなリードI/Oを選び分ける精度が低下するという点である。この点は、シーケンシャルなリードI/Oのキャッシュヒット率を低下させることになる。

【0029】2点目は、複数のキャッシュ制御情報に共に断続的なアドレスのリードI/Oが記録されるため、複数のホスト制御部にてシーケンシャルなリードI/Oだと判断してしまい、同一のアドレス領域へのプリフェッチを多重に行なってしまう可能性がある点である。この点は、そのアドレス領域に属する磁気ディスクの混雑を招くことになる。すなわち、シーケンシャルなリードI/Oのレスポンスタイムを大幅に悪化させてしま

う。

【0030】上記した例は、負荷分散アルゴリズムに、ラウンドロビン以外の特開平07-334449号公報に開示されたような方法を用いた場合でも同様である。これは、非共有キャッシュ方式のデータ記憶装置に対してマルチバス負荷分散を用いると、複数のキャッシュ制御情報にそれぞれ断続的なアドレスへのリードI/Oが記録されることになるからである。

【0031】本発明の目的は、上記従来技術の欠点を解決し、非共有キャッシュ方式のデータ記憶装置にアクセスするホストコンピュータが、データ記憶装置のホスト制御部によって行なわれるプリフェッチの性能を低下させることなくマルチバス負荷分散を行なうことができる負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラムを提供することにある。

【0032】特に、ホスト制御部がシーケンシャルなリードI/Oを選り分ける精度が低下する問題、複数のホスト制御部から同一のアドレス領域へのプリフェッチを多重に行なってしまう問題を発生させずにマルチバス負荷分散を行なうことができる負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラムを提供することにある。

【0033】

【課題を解決するための手段】上記目的を達成するため本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする。

【0034】請求項2の本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理し、前記バスグループの情報を示すバスグループ情報を生成し、入出力要求のアドレス情報と前記バスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする。

【0035】請求項3の本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散シ

テムにおいて、前記データ記憶装置と前記ホストコンピュータとの間の通信を中継し、前記データ記憶装置のキャッシングを実行するホスト制御装置を備え、入出力要求のアドレス情報と、前記ホストコンピュータと前記ホスト制御装置との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする。

【0036】請求項4の本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、前記データ記憶装置と前記ホストコンピュータとの間の通信を中継し、前記データ記憶装置のキャッシングを実行するホスト制御装置を備え、前記ホストコンピュータと前記ホスト制御装置との間における複数の物理的バスを、前記ホスト制御装置内のキャッシュ毎にグループ化してバスグループとして管理し、前記バスグループの情報を示すバスグループ情報を生成し、入出力要求のアドレス情報と前記バスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする。

【0037】請求項5の本発明の負荷分散システムは、前記マッピング情報の生成において、前記入出力要求のアドレス情報として、当該入出力要求の論理ボリューム番号を用いることを特徴とする。

【0038】請求項6の本発明の負荷分散システムは、前記入出力要求の論理アドレス番号を適当な幅毎に区画化し、前記マッピング情報の生成において、前記入出力要求のアドレス情報として、当該入出力要求の論理アドレス番号が該当する前記区画を用いることを特徴とする。

【0039】請求項7の本発明の負荷分散システムは、前記ネットワークを介して、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする。

【0040】請求項8の本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、前記ホストコンピュータが、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理するバスグループ情報を記録するバスグループテーブルと、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングした、入出力要求のアドレス情報と前記バスグループとのマッピングを示す

マッピング情報を記録するマッピングテーブルを備え、前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散することを特徴とする。

【0041】請求項9の本発明のホストコンピュータは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理手段と、前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御手段を備えることを特徴とする。

【0042】請求項10の本発明のホストコンピュータは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理し、前記バスグループの情報を示すバスグループ情報を生成するバスグループ管理手段と、入出力要求のアドレス情報と前記バスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理手段と、前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御手段を備えることを特徴とする。

【0043】請求項11の本発明のホストコンピュータは、前記バスグループ管理手段は、前記データ記憶装置から、当該データ記憶装置が記憶するデータの構成情報を取得する手段を備えることを特徴とする。

【0044】請求項12の本発明のホストコンピュータは、前記バスグループ管理手段は、指定された外部の情報機器から、前記データ記憶装置が記憶するデータの構成情報を取得する手段を備えることを特徴とする。

【0045】請求項13の本発明のホストコンピュータは、前記マッピング管理手段は、前記入出力要求のアドレス情報として、当該入出力要求の論理ボリューム番号を用いてマッピングを行なうことを特徴とする。

【0046】請求項14の本発明のホストコンピュータは、前記マッピング管理手段は、前記入出力要求の論理アドレス番号を適当な幅毎に区画化し、前記入出力要求のアドレス情報として、当該入出力要求の論理アドレス番号が該当する前記区画を用いてマッピングを行なうことを特徴とする。

【0047】請求項15の本発明のホストコンピュータは、前記マッピング管理手段は、前記データ記憶装置か

ら、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする。

【0048】請求項16の本発明のホストコンピュータは、前記マッピング管理手段は、指定された外部の情報機器から、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする。

【0049】請求項17の本発明のホストコンピュータは、前記外部の情報機器を、ネームサーバとすることを特徴とする。

【0050】請求項18の本発明のホストコンピュータは、前記マッピング管理手段は、定期的に各前記バスの負荷情報を参照して前記マッピング情報を更新することを特徴とする。

【0051】請求項19の本発明のホストコンピュータは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理するバスグループ情報を記録するバスグループテーブルと、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングした、入出力要求のアドレス情報と前記バスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルと、前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御手段を備えることを特徴とする。

【0052】請求項20の本発明の負荷分散プログラムは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理機能と、前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御機能を備えることを特徴とする。

【0053】請求項21の本発明の負荷分散プログラムは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理し、前記バスグループの情報を示すバスグループ情報を生成するバスグループ管理機能と、入出力要求のアドレス情報と前記バスグル

ープとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理機能と、前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御機能を備えることを特徴とする。

【0054】請求項22の本発明の負荷分散プログラムは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、前記ホストコンピュータが、前記データ記憶装置との間における複数の物理的バスを、前記データ記憶装置内のキャッシュ毎にグループ化してバスグループとして管理するバスグループ情報を記録するバスグループテーブルと、シーケンシャルな読み出し要求を単一の前記バスグループにマッピングした、入出力要求のアドレス情報と前記バスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルを備え、前記バスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的バスに適切に分散する負荷分散制御機能を備えることを特徴とする。

【0055】本発明では、バスグループ管理部とマッピング管理部とにより、シーケンシャルな読み出し要求を単一のキャッシュに送付する制限を加える。このような制限を加えることで、シーケンシャルな読み出し要求は、キャッシュを共有しているホスト制御部群によって処理されるようになる。すなわち、その上で負荷分散を行なうことにより、ホスト制御部がシーケンシャルな読み出し要求を選り分ける精度が低下する問題、複数のホスト制御部から同一のアドレス領域へのプリフェッチを多重に行なってしまう問題を解消することができる。

【0056】また、本発明の第2の実施の形態では、バスグループの情報と、入出力要求のアドレス情報とバスグループとのマッピング情報を、管理者が前もってバスグループテーブル、マッピングテーブルとしてホストコンピュータに与えておくことにより、負荷分散制御部がこれらのテーブルを参照し、負荷分散アルゴリズムに基づき使用するバスを決定する。

【0057】

【発明の実施の形態】以下、本発明の実施の形態について図面を参照して詳細に説明する。

【0058】図1は、本発明の第1の実施の形態による負荷分散システムの構成を示すブロック図である。図1を参照すると、本発明の第1の実施の形態は、プログラム制御により動作するホストコンピュータ100と、非共有キャッシュ方式のデータ記憶装置200、ネットワーク300から構成される。

【0059】ホストコンピュータ100は、アプリケーション110と、ファイルシステム120と、負荷分散

制御部130と、負荷分散情報テーブル131と、バスグループ管理部140と、バスグループテーブル141と、マッピング管理部150と、マッピングテーブル151と、デバイスドライバ160～165と、通信ポート170～175を含む。

【0060】データ記憶装置200は、ホスト制御部210、211と、キャッシュ220、221と、通信ポート230～235と、ディスクアレイ240を含む。

【0061】なお、本形態ではアプリケーション110、ファイルシステム120は、簡便のため1つのみ示したが、複数のアプリケーション、複数のファイルシステムが単一のホストコンピュータ上で動作している場合も考えられる。

【0062】またデバイスドライバ160～165、通信ポート170～175、230～235（ホストコンピュータ側、データ記憶装置側双方）は、簡便のため6つずつ示したが、それぞれ2～50程度備えている場合も考えられる。

【0063】データ記憶装置200は、簡便のため1つのみ示したが、複数のデータ記憶装置と接続されている場合も考えられる。

【0064】ホスト制御部210、211は、簡便のため2つのみ示したが、2～20程度備えている場合も考えられる。また、各ホスト制御部210、211は3つずつ通信ポートを持っているが、1～10程度備えている場合も考えられるし、ホスト制御部間210、211で異なる数の通信ポートを持つ場合や、異なる性能の通信ポートが混在する場合も考えられる。また、この図1では、1つのホスト制御部が1つのキャッシュを持つ構成になっているが、複数のホスト制御部が1つのキャッシュを共有する場合も考えられる。

【0065】また、ネットワーク300は、いろいろな形態が考慮できる。具体的な例としてはFiber Channel Arbitrated Loop、Fiber Channel Fabric、Ethernet、Parallel SCSI (Small Computer System Interface)、ATM (Asynchronous Transfer Mode)、FDDI (Fiber Distributed Data Interface) 等のデータリンクから構成されることが考えられる。また、途中経路にリピーター、ブリッジ、ルーター、ゲートウェイ等の相互接続機器が設けられていることも考えられる。

【0066】これらの各部は、それぞれ概略つぎのように動作する。

【0067】ファイルシステム120は、アプリケーション110からI/O要求（入出力要求）を受け取ると、それらをブロック単位のI/O要求に変換し、負荷分散制御部130に送付する。各I/O要求には、LUN、LBA等のアドレス情報が付加される。

【0068】負荷分散制御部130は、ファイルシステム120からI/O要求を受け取ると、まずI/O要求のアドレス情報とマッピングテーブル151を参照して

使用するバスグループを決定する。

【0069】そして、バスグループテーブル141を参照してバスグループに属するバスを確認し、その中から負荷分散アルゴリズムに基づき使用するバスを決定する。最後に、デバイスドライバ160～165の中の適切なものに対して、I/O要求を送付する。

【0070】バスグループ管理部140は、まず各データ記憶装置に対するバスを列挙し、そして各バスが接続されているホスト制御部がデータ記憶装置のどのキャッシュを使用しているかを確認する。そして、使用しているキャッシュ毎にバスをグループ化し、その情報をバスグループテーブル141に記述する。

【0071】マッピング管理部150は、まずファイルシステム120が担当するI/O要求のアドレス情報の範囲を特定し、アドレス情報とバスグループのマッピングを行なう。そして、その情報をマッピングテーブル151に記述する。

【0072】データ記憶装置200は、複数のキャッシュ220、221を備えており、非共有キャッシュ方式を採用している。ダーティデータに関しては、分散されたキャッシュ220、221間でコピーを行なうことで、キャッシュ間でのキャッシュの一貫性を維持している。それ以外のキャッシングデータに関してはコピーを行わないので、キャッシングデータの内容はキャッシュ間によって異なる。

【0073】ホスト制御部210、211は、通信ポート230～235からI/O要求を受け取ると、そのアドレス情報を参照して、ディスクアレイ240に対してI/O要求を送付する。またキャッシュ220、221を用いてリードキャッシング、ライトキャッシング、プリフェッチを行なう。

【0074】ディスクアレイ240は、複数の磁気ディスク装置を備え、ホスト制御部210、211からI/O要求を受け取ると、そのアドレス情報を参照して適切な磁気ディスクへI/O要求を送付する。また、内部に複数のRAIDコントローラを備え、RAID装置として動作することも考えられる。

【0075】次に、図2～図4のフローチャートを参照して本実施の形態の全体の動作について詳細に説明する。

【0076】図2は、本実施の形態の負荷分散制御部130の動作を説明するためのフローチャートである。負荷分散制御部130は、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。そして、ファイルシステム120よりI/O要求が来るまでは待機している(ステップ201)。

【0077】ファイルシステム120よりI/O要求が来ると、そのI/O要求のアドレス情報とマッピングテーブル151を参照して使用するバスグループを選択する(ステップ202)。マッピングテーブル151に

は、I/O要求のアドレス情報であるLUNやLBAからバスグループが一意に選択できるようなマッピング情報が記述されている。

【0078】次に、バスグループテーブル141、負荷分散情報テーブル131を参照して使用するバスを選択する(ステップ203)。バスグループテーブル141には、各バスグループにはどのバスが属するかが記述されており、これを参照することでバスグループに属するバスを確認できる。

【0079】負荷分散情報テーブル131には、各バスの利用状況と、各バスグループにおいてどの負荷分散アルゴリズムを用いるかということが記述されており、バスグループテーブル141の情報と組み合わせて使用するバスを選択することができる。

【0080】ここでいうバスの利用状況とは、各バスグループの中で直前にI/O要求を送付したバス、各バスにてペンディングされているI/O要求の数、各バスにてペンディングされているI/O要求のブロック数の総和等を指す。

【0081】また、ここでいうバスグループの負荷分散アルゴリズムとは、ラウンドロビンにて送付する方法、ペンディングされているI/O要求の数が最も少ないバスに送付する方法、ペンディングされているI/O要求のブロック数の総和が最も少ないバスに送付する方法等を指す。

【0082】例えば、負荷分散にラウンドロビン方式を用いている場合の具体的な動作手順を以下に示す。まず、ステップ202で選択したバスグループにて用いられる負荷分散アルゴリズムを、負荷分散情報テーブル131を参照して調べる(この例ではラウンドロビン方式だという結果が出る)。次に、バスグループテーブル141を参照して選択したバスグループに属するバスを調べる。最後に、負荷分散情報テーブル131内のバスの利用状況を参照し使用するバスが決定する(この例では直前にI/O要求を送付したバスの情報を参照することで、次のバスを決定できる)。

【0083】この手順は、異なる負荷分散アルゴリズムを用いている場合でもほぼ同様である。ただ異なる利用状況に関する情報を参照するだけである。

【0084】使用するバスが決定されると、適切なデバイスドライバへI/O要求を送付する(ステップ204)。最後に、新しくI/O要求を送付したことで変化したバスの利用状況を、負荷分散情報テーブル131に記述する(ステップ205)。この後は、再びファイルシステム120よりI/O要求が来るまでは、待機することになる。

【0085】図3は、本実施の形態のバスグループ管理部140の動作を説明するためのフローチャートである。バスグループ管理部140は、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。

そして、各データ記憶装置に対するパスを列挙する（ステップ301）。

【0086】パスを列挙する分解能は、エンドツーエンドの通信ポートの組合せ、すなわちホストコンピュータ側の各通信ポートと、そこからアクセス可能なデータ記憶装置側の通信ポートの組合せになる。例えば、ホストコンピュータ側の6個の通信ポートがそれぞれ2個のデータ記憶装置側の通信ポートにアクセス可能ならば、全部で12個のパスが列挙される。

【0087】アクセス可能であるという意味は、物理的に可能であるという意味だけでなく、ゾーニング等のアクセスコントロールによる制約も受けていないという意味も含んでいる。つまり、このステップ301では、アクセス可能なエンドツーエンドのパスを全て列挙することである。

【0088】パスが列挙されたなら、各パスが接続されているホスト制御部がデータ記憶装置のどのキャッシュを使用しているかを確認する（ステップ302）。このステップ302では、まず各パスがデータ記憶装置側のどの通信ポートを用いているか特定する。そして、その通信ポートを使っているホスト制御部がどのキャッシュを使っているか調べる。

【0089】例えば、本実施の形態では、通信ポート230～232を用いている場合はキャッシュ220を使用し、通信ポート233～235を用いている場合はキャッシュ221を使用しているといったように判別する。

【0090】ステップ302では、データ記憶装置の構成情報、すなわち、通信ポートと使用しているキャッシュの対応関係の情報をどうやって知るかで2つの方法に分かれる。つまり、データ記憶装置に教えてもらう方法と、ネームサーバ等の第三者に教えてもらう方法とがある。

【0091】1つ目の方法は、例えば各パスがデータ記憶装置のどの通信ポートを用いているか特定する際に、ホストコンピュータとデータ記憶装置間で通信が発生するから、その際にデータ記憶装置側からホストコンピュータへデータ記憶装置の構成情報を渡すというものである。

【0092】2つ目は、アクセスしたいデータ記憶装置側の通信ポートを特定する際にネームサーバに問い合わせる場合、ネームサーバに対応情報を与えておくことで、ホストコンピュータは、通信ポートと使用するキャッシュとの対応関係を、ネームサーバを通して知るというものである。

【0093】各パスがどのキャッシュを使用しているかを確認したら、使用しているキャッシュ毎にパスをグループ化する（ステップ303）。このグループ化したパス群はパスグループとして扱われることになる。最後に、パスグループの情報をパスグループテーブル141

に記述する（ステップ304）。

【0094】図4は、本実施の形態のマッピング管理部150の動作を説明するためのフローチャートである。マッピング管理部150は、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。そして、まずファイルシステム120が担当するI/O要求のアドレス情報の範囲を特定する（ステップ401）。具体的には、LUNやLBAが使用されている範囲が何番から何番までかを特定するということである。

【0095】次に、アドレス情報とパスグループのマッピングを行なう（ステップ402）。ステップ402のマッピングの目的は、シーケンシャルなリードI/Oを単一のパスグループに導きつつ、ランダムなI/Oはなるべくそのような制限を設けずに、従来と同様な負荷分散アルゴリズムを用いてマルチパス負荷分散を行なうことができるようにすることにある。具体的にはI/O要求のアドレス情報であるLUNとLBAから、一意のパスグループが導かれるように行なう。以下に例を2つ示す。

【0096】1つ目は論理ボリューム番号（以下LUN）を用いる方法である。この場合は、各LUNを適当に（例えば昇順に）パスグループに対してマッピングする。この方法は、シーケンシャルなI/OはほとんどLUNが同じであるという特性を利用している。すなわち、シーケンシャルなI/Oはほとんど同じパスグループにマッピングされることになる。

【0097】2つ目は論理アドレス（以下LBA）の区画を用いる方法である。この場合は、LBAを或る適当な幅毎に区画化し、各区画を適当に（例えば昇順に）パスグループに対してマッピングしていく。この方法は、シーケンシャルなI/OはLBAがかなり接近しているという特性を利用している。すなわち、シーケンシャルなI/Oはほとんど同じパスグループにマッピングされることになる。

【0098】また、ここではいずれか1つを用いる方法について説明したが、LUNとLBAを組み合わせるパスグループを導く方法も考えられる。

【0099】ステップ402においてLBAの区画を用いる場合は、その区画幅の決め方として2つ方法が考えられる。つまり、データ記憶装置に教えてもらう方法と、ネームサーバ等の第三者に教えてもらう方法とがある。

【0100】1つ目の方法は、例えば各パスがデータ記憶装置のどの通信ポートを用いているか特定する際に、ホストコンピュータとデータ記憶装置間で通信が発生するから、その際にデータ記憶装置側からホストコンピュータへ最適な区画幅を渡すというものである。

【0101】2つ目は、ホストコンピュータはネームサーバを通して最適な区画幅の情報を知るというものである。

【0102】アドレス情報とバスグループのマッピングを行なったら、その情報をマッピングテーブル151に記述する(ステップ403)。

【0103】次に、本実施の形態の効果について説明する。以上説明したように、本実施の形態によれば、バスグループ管理部140、マッピング管理部150は合わせて、シーケンシャルなリードI/Oを単一のキャッシュに送付する制限を加えるための情報を提供している。このような制限を加えることで、シーケンシャルなリードI/Oは、キャッシュを共有しているホスト制御部群(210、211)によって処理されるようになる。そして、その上で負荷分散制御部130が負荷分散を行なっているので、ホスト制御部210、211がシーケンシャルなリードI/Oを選び分ける精度が低下する問題、複数のホスト制御部210、211から同一のアドレス領域へのプリフェッチを多重に行なってしまう問題は発生しない。

【0104】次に、本発明の第2の実施の形態について図面を参照して詳細に説明する。図5は、本発明の第2の実施の形態による負荷分散システムの構成を示すブロック図である。

【0105】図5を参照すると、本発明の実施の形態の負荷分散システムの構成の第1の実施の形態との相違点は、ホストコンピュータ100aにおいて、第1の実施の形態におけるバスグループ管理部140とマッピング管理部150を備えない点である。

【0106】すなわち、接続されているホスト制御部210、211が使用するデータ記憶装置200のキャッシュ毎にバスをグループ化したバスグループの情報と、I/O要求のアドレス情報とバスグループのマッピング情報を、管理者が前もってバスグループテーブル141a、マッピングテーブル151aとしてホストコンピュータ100aに与えておく。

【0107】負荷分散制御部130aは、これらのテーブルを参照し、負荷分散アルゴリズムに基づき使用するバスを決定する。

【0108】以上説明したように本実施の形態によれば、第1の実施の形態の効果に加えて、情報を人間が設定するために非常に細やかな設定が可能となる点や、システムを非常にシンプルに構成することができるという点がある。

【0109】次に、本発明の第3の実施の形態について詳細に説明する。図6は、本発明の第3の実施の形態による負荷分散システムの構成を示すブロック図である。

【0110】本発明の第3の実施の形態の負荷分散システムの構成は、第1の実施の形態と同様であるが、ホストコンピュータ100bのマッピング管理部150bにおいて、定期的にマッピングテーブル151を更新する機能を備えるようにした点を特徴とする。

【0111】図7は、本実施の形態のマッピング管理部

150bの動作を説明するためのフローチャートである。図7を参照すると本実施の形態のマッピング管理部150bは、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。そして、まずファイルシステム120が担当するI/O要求のアドレス情報の範囲を特定する(ステップ701)。

【0112】次に、管理者に定められた時刻が来るまで待つ(ステップ702)。定められた時刻に達すると、アドレス情報とバスグループのマッピングを行なう(ステップ703)。アドレス情報とバスグループのマッピングを行なったら、その情報をマッピングテーブル151に記述する(ステップ704)。この後は、再び管理者に定められた時刻が来るまでは、待機することになる。

【0113】第1の実施の形態と異なる点は、マッピング管理部150bが、定期的にマッピングテーブル151を、負荷分散情報テーブル131を参照して更新する点である。例えば、本実施の形態のマッピング管理部150bは、毎日午前12時にその一日のバスの利用状況を参照してマッピングテーブルを更新するような形になる。

【0114】定期的にマッピングテーブル151を、負荷情報を参照して更新することで、バスの負荷状況に沿ってバスグループとI/Oのアドレス情報をマッピングすることが可能になる。引いては、より適切なマルチバス負荷分散が可能になる。

【0115】以上説明したように本実施の形態によれば、第1の実施の形態の効果に加えて、定期的に負荷情報を参照してマッピングテーブル151を更新することで、バスの負荷状況に沿ってバスグループとI/Oのアドレス情報をマッピングすることが可能になる。バスの負荷状況に沿ったマッピングは、より適切なマルチバス負荷分散を実現する。

【0116】次に、本発明の第4の実施の形態について詳細に説明する。

【0117】図8は、本発明の第4の実施の形態による負荷分散システムの構成を示すブロック図である。図8を参照すると、本実施の形態の負荷分散システムの構成の第1の実施の形態との相違点は、ホストコンピュータ100cにおいて、バスグループ管理部140とバスグループテーブル141を備えない点であり、また、マッピング管理部150cの機能を変更することにより、バスグループという概念を本実施の形態においては取り扱わないことを特徴とする。

【0118】本実施の形態のマッピング管理部150cは、I/Oのアドレス情報とバスを直接マッピングする。そして、負荷分散制御部130cは、そのマッピング情報を用いることで、シーケンシャルなリードI/Oを単一のバスに導きつつ、ランダムなI/Oはなるべくそのような制限を設けずにマルチバス負荷分散を行な

う。

【0119】図9は、本発明の第4の実施の形態の負荷分散制御部130cの動作を示すフローチャートである。負荷分散制御部130cは、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。そして、ファイルシステム120よりI/O要求が来るまでは待機している(ステップ901)。

【0120】ファイルシステム120よりI/O要求が来ると、そのI/O要求のアドレス情報とマッピングテーブル151と負荷分散情報テーブル131を参照して使用するバスを選択する(ステップ902)。マッピングテーブル151には、I/O要求のアドレス情報であるLUNやLBAからバスが一意に選択できるようなマッピング情報が記述されている。負荷分散情報テーブル131には、各バスの利用状況と、各バスグループにおいてどの負荷分散アルゴリズムを用いるかということが記述されている。

【0121】使用するバスが決定されると、適切なデバイスドライバへI/O要求を送付する(ステップ903)。最後に、新しくI/O要求を送付したことで変化したバスの利用状況を負荷分散情報テーブル131に記述する(ステップ904)。この後は再びファイルシステム120よりI/O要求が来るまでは待機することになる。

【0122】図10は、本発明の第4の実施の形態のマッピング管理部150cの動作を表すフローチャートである。マッピング管理部150cは、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。

【0123】そして、まずファイルシステム120が担当するI/O要求のアドレス情報の範囲を特定する(ステップ1001)。具体的には、LUNやLBAが使用されている範囲が何番から何番までかを特定することである。

【0124】次に、アドレス情報とバスのマッピングを行なう(ステップ1002)。ステップ1002のマッピングの目的は、シーケンシャルなリードI/Oを単一のバスに導きつつ、ランダムなI/Oはなるべくそのような制限を設けずに、従来と同様な負荷分散アルゴリズムを用いてマルチバス負荷分散を行なうことができるようにすることにある。具体的には、I/O要求のアドレス情報であるLUNとLBAから、一意のバスが導かれるように行なう。

【0125】アドレス情報とバスのマッピングを行なったら、その情報をマッピングテーブル151に記述する(ステップ1003)。

【0126】以上説明したように本実施の形態によれば、第1の実施の形態の効果に加えて、バスグループという概念を必要としないことから、システムをシンプルに構成できる。

【0127】次に、本発明の第5の実施の形態について詳細に説明する。

【0128】図11は、本発明の第5の実施の形態による負荷分散システムの構成を示すブロック図である。図11を参照すると、本発明の第5の実施の形態は、ホストコンピュータ100と、非共有キャッシュ方式のホスト制御装置400、データ記憶装置200d、ネットワーク300を備える。

【0129】図11を参照すると、本実施の形態の負荷分散システムの構成の第1の実施の形態との相違点は、仮想ボリューム機能やキャッシング機能を行なうホスト制御部410、411を擁するホスト制御装置400を、ディスクアレイ240を備えるデータ記憶装置200dと独立させた点である。本実施の形態において、ホスト制御装置400において、データ記憶装置200dの仮想ボリューム機能やキャッシング機能を提供する。このため、データ記憶装置200d自身には、ホスト制御部を備える必要はない。

【0130】本実施の形態のホストコンピュータ100の処理は、第1の実施の形態と同様である。ただし、本実施の形態においてマッピングを行なう物理的バスは、ホストコンピュータ100とホスト制御装置400との間における物理的バスである。

【0131】つまり、ホストコンピュータ100とホスト制御装置400との間における複数の物理的バスを、ホスト制御装置400内のキャッシュ毎にグループ化してバスグループとして管理し、バスグループの情報を示すバスグループ情報を生成する。そして、入出力要求のアドレス情報とバスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一のバスグループにマッピングし、マッピングの内容を示すマッピング情報を生成する。そして、生成されたマッピング情報を参照して、第1の実施の形態において説明されたように、入出力要求を複数本の物理的バスに適切に分散する。

【0132】なお、図11の例においては、ホスト制御装置400は簡便のため1つのみ示したが、複数のホスト制御装置が連携して動作している場合も考えられる。

【0133】ホスト制御部410、411は、簡便のため2つのみ示したが、2~20程度備えている場合も考えられる。また、各ホスト制御部410、411は3つずつ通信ポートを持っているが、1~10程度備えている場合も考えられるし、ホスト制御部間410、411で異なる数の通信ポートを持つ場合や、異なる性能の通信ポートが混在する場合も考えられる。また、この図11では、1つのホスト制御部が1つのキャッシュを持つ構成になっているが、複数のホスト制御部が1つのキャッシュを共有する場合も考えられる。

【0134】またホスト制御装置400は、ディスクアレイ、ハードディスク、半導体メモリ等のデータ記憶装置を持つ場合も考えられる。



【0135】またデータ記憶装置200dは、簡便のためディスクアレイのみ示したが、ホスト制御部、キャッシュを持つ場合も考えられる。また、データ記憶装置200dは、簡便のため通信ポートを1つのみ示したが、2～50程度備えている場合も考えられる。

【0136】本実施の形態によれば、第1の実施の形態の効果に加えて、スト制御部を擁するホスト制御装置とディスクアレイを備えるデータ記憶装置とを独立の装置することができ、データ記憶装置の構成には手を加えずに、ホスト制御装置を加えたり削除したりする等の変更が容易になるため、ホスト制御部が実現する仮想ボリューム機能、キャッシング機能の性能をより柔軟に調整することが可能になる。

【0137】また、上記各実施の形態の負荷分散システムは、互いに自由に組み合わせて実施することができる。例えば、第5の実施の形態において、第2、3、4の実施の形態のホストコンピュータを用いて実施すること等が可能である。

【0138】例えば、第5の実施の形態において、第4の実施の形態のホストコンピュータ100cを用いる場合には、入出力要求のアドレス情報と、ホストコンピュータ100cとホスト制御装置400との間における複数の物理的バスとのマッピングを行ない、シーケンシャルな読み出し要求を単一のバスにマッピングし、マッピングの内容を示すマッピング情報を生成する。そして、生成されたマッピング情報を参照して、入出力要求を複数本の物理的バスに適切に分散する。

【0139】なお、上記各実施の形態の負荷分散システムは、ホストコンピュータ100、100a、100b、100cにおける負荷分散制御部130、130a、130c、バスグループ管理部140、マッピング管理部150、150b、150cや、その他の機能をハードウェア的に実現することは勿論として、各機能を備えるコンピュータプログラムである負荷分散プログラム90、90a、90b、90c、90dを、コンピュータ処理装置のメモリにロードされることで実現することができる。この負荷分散プログラム90、90a、90b、90c、90dは、磁気ディスク、半導体メモリその他の記録媒体に格納される。そして、その記録媒体からコンピュータ処理装置にロードされ、コンピュータ処理装置の動作を制御することにより、上述した各機能を実現する。

【0140】以上好ましい実施の形態及び実施例をあげて本発明を説明したが、本発明は必ずしも上記実施の形態及び実施例に限定されるものではなく、その技術的思想の範囲内において様々に変形して実施することができる。

【0141】

【発明の効果】以上説明したように本発明の負荷分散システム、負荷分散システムのホストコンピュータ、及び

負荷分散プログラムによれば、以下のような効果が達成される。

【0142】第1に、本発明によれば、各バスが接続されているホスト制御部がデータ記憶装置のどのキャッシュを使用しているかを確認し、使用しているキャッシュ毎にバスをグループ化するバスグループ管理部と、I/O要求（入出力要求）のアドレス情報とバスグループのマッピングを行なうマッピング管理部により、シーケンシャルなリードI/Oを単一のキャッシュに送付する制限を加えた上でマルチバス負荷分散を行なうことができる。すなわち、ホスト制御部がシーケンシャルなリードI/Oを選り分ける精度が低下する問題、複数のホスト制御部から同一のアドレス領域へのプリフェッチを多重に行なってしまう問題は発生しない。

【0143】第2に、本発明の第2の実施の形態によれば、バスグループの情報と、I/O要求のアドレス情報とバスグループのマッピング情報を、管理者が前もってバスグループテーブル、マッピングテーブルとしてホストコンピュータに与えておくことで、システムをシンプルに構成することができる。

【0144】第3に、本発明の第3の実施の形態によれば、定期的に負荷情報を参照してI/O要求のアドレス情報とバスグループのマッピングを更新することで、バスの負荷状況に沿ってバスグループとI/Oのアドレス情報をマッピングすることができる。バスの負荷状況に沿ったマッピングは、より適切なマルチバス負荷分散を実現する。

【0145】第4に、本発明の第4の実施の形態によれば、バスグループという概念を用いずに、I/O要求のアドレス情報とバスを直接マッピングすることで、システムをシンプルに構成することができる。

【図面の簡単な説明】

【図1】 本発明の第1の実施の形態による負荷分散システムの構成を示すブロック図である。

【図2】 本発明の第1の実施の形態の負荷分散制御部の動作を説明するためのフローチャートである。

【図3】 本発明の第1の実施の形態のバスグループ管理部の動作を説明するためのフローチャートである。

【図4】 本発明の第1の実施の形態のマッピング管理部の動作を説明するためのフローチャートである。

【図5】 本発明の第2の実施の形態による負荷分散システムの構成を示すブロック図である。

【図6】 本発明の第3の実施の形態による負荷分散システムの構成を示すブロック図である。

【図7】 本発明の第3の実施の形態のマッピング管理部の動作を説明するためのフローチャートである。

【図8】 本発明の第4の実施の形態による負荷分散システムの構成を示すブロック図である。

【図9】 本発明の第4の実施の形態の負荷分散制御部の動作を説明するためのフローチャートである。

【図10】 本発明の第4の実施の形態のマッピング管理部の動作を説明するためのフローチャートである。

【図11】 本発明の第5の実施の形態による負分散システムの構成を示すブロック図である。

【図12】 従来の負分散システムの構成を示すブロック図である。

【図13】 従来の共有キャッシュ方式と非共有キャッシュ方式のそれぞれのデータ記憶装置の構成を示すブロック図である。

【符号の説明】

90、90a、90b、90c、90d 負分散プログラム

100、100a、100b、100c ホストコンピュータ

110 アプリケーション

120 ファイルシステム

130、130a、130c 負分散制御部

131 負分散情報テーブル

140 バスグループ管理部

141、141c バスグループテーブル

150、150b、150c マッピング管理部

151、151a マッピングテーブル

160～165 デバイスドライバ

170～175 通信ポート

200、200d データ記憶装置

210、211、410、411 ホスト制御部

220、221、220、221 キャッシュ

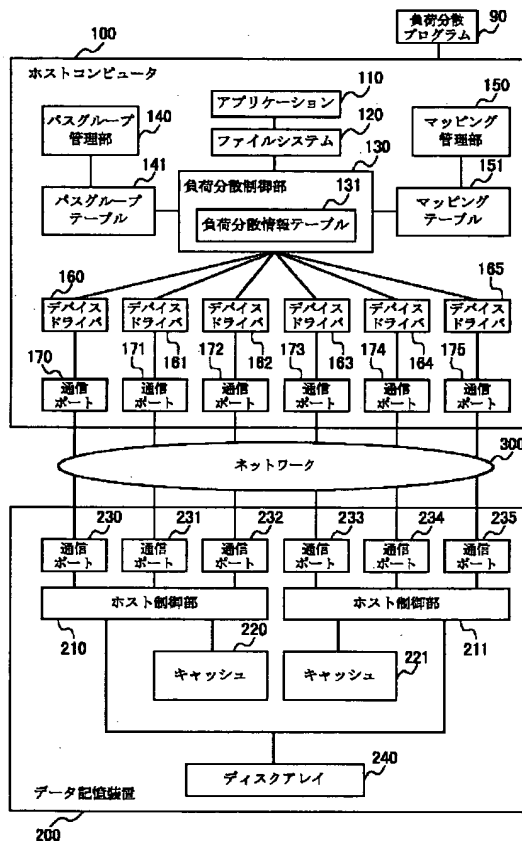
230～235、430～435 通信ポート

240 ディスクアレイ

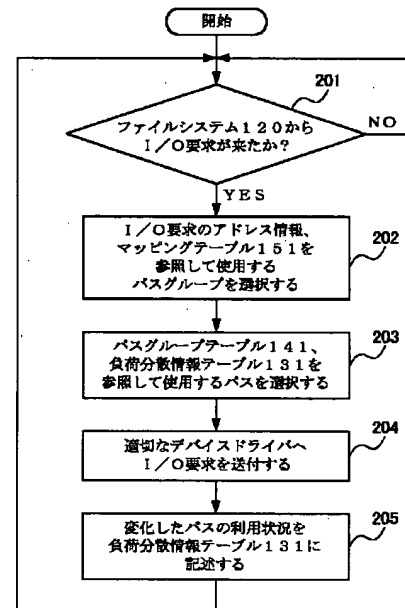
300 ネットワーク

400 ホスト制御装置

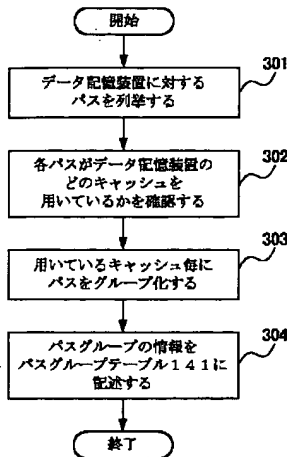
【図1】



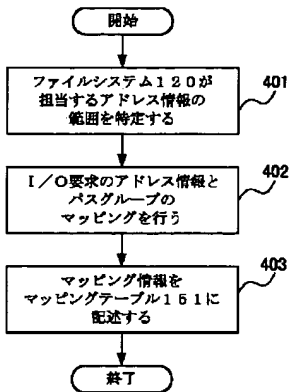
【図2】



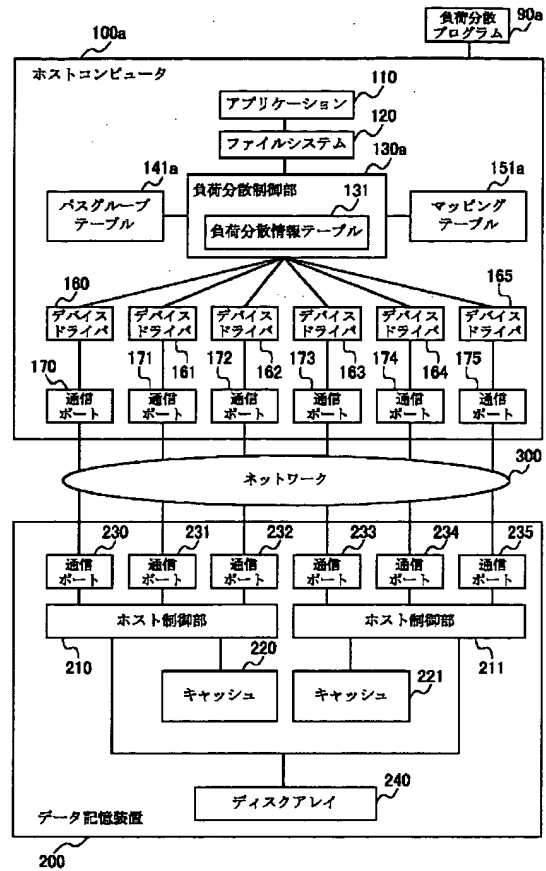
【図3】



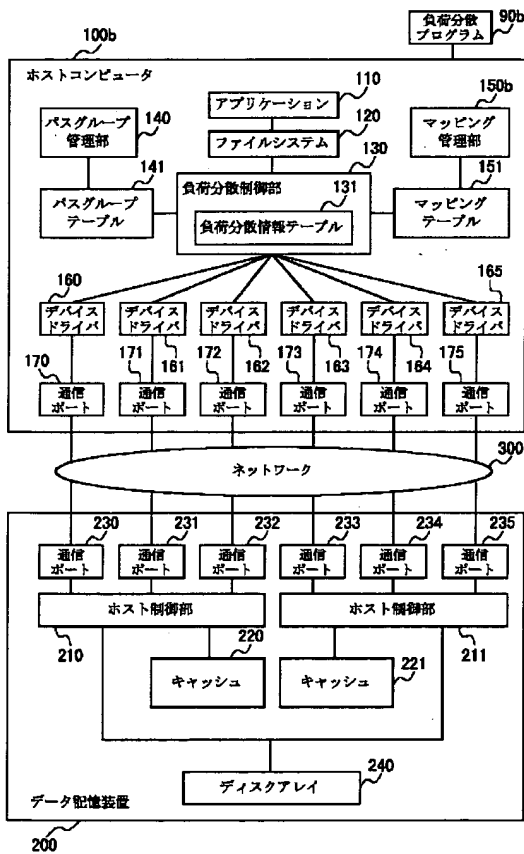
【図4】



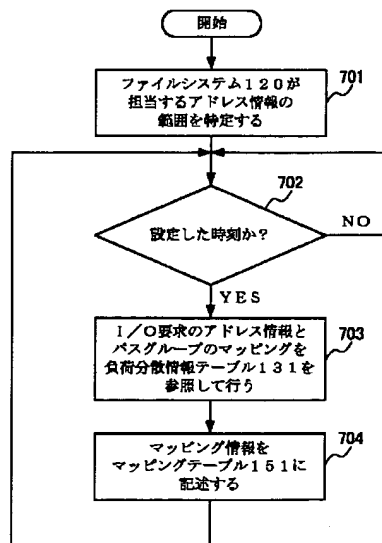
【図5】



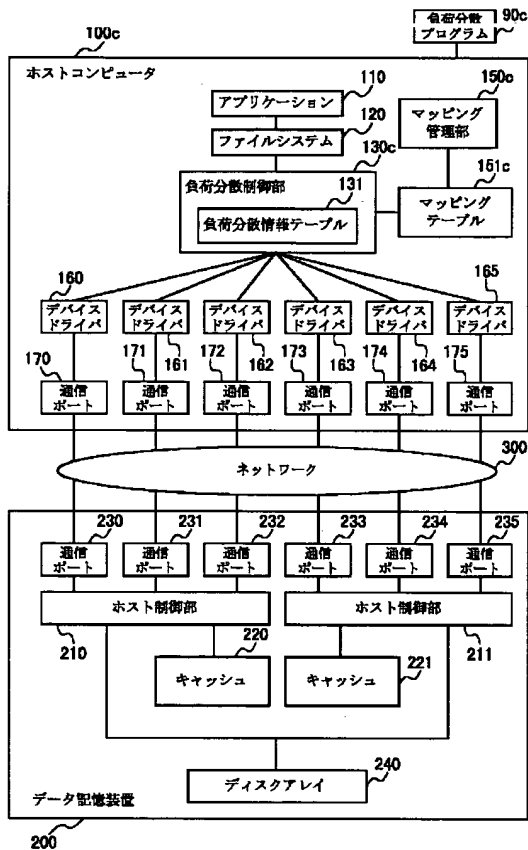
【図6】



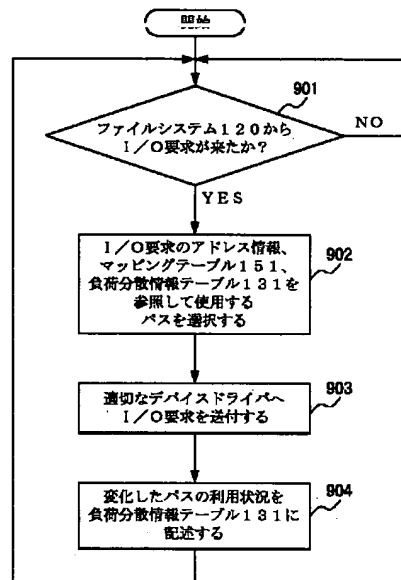
【図7】



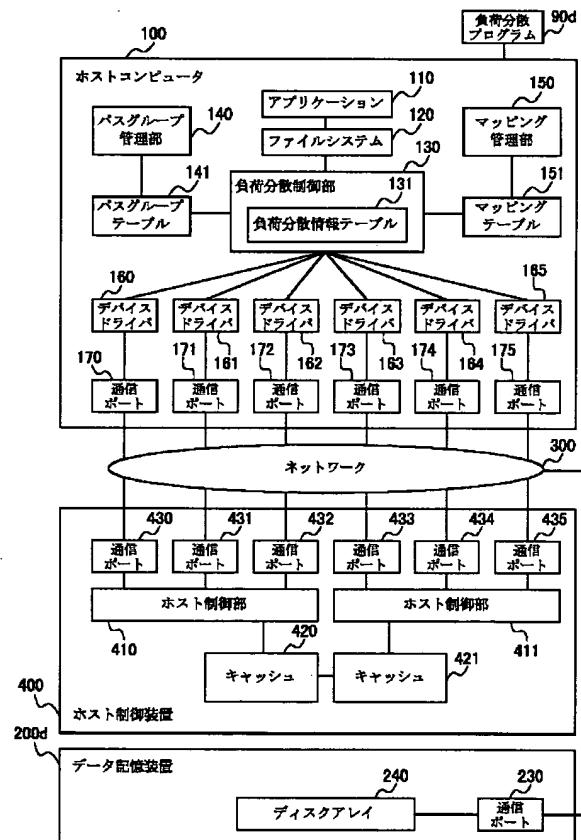
【図8】



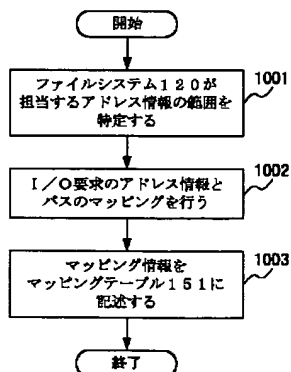
【図9】



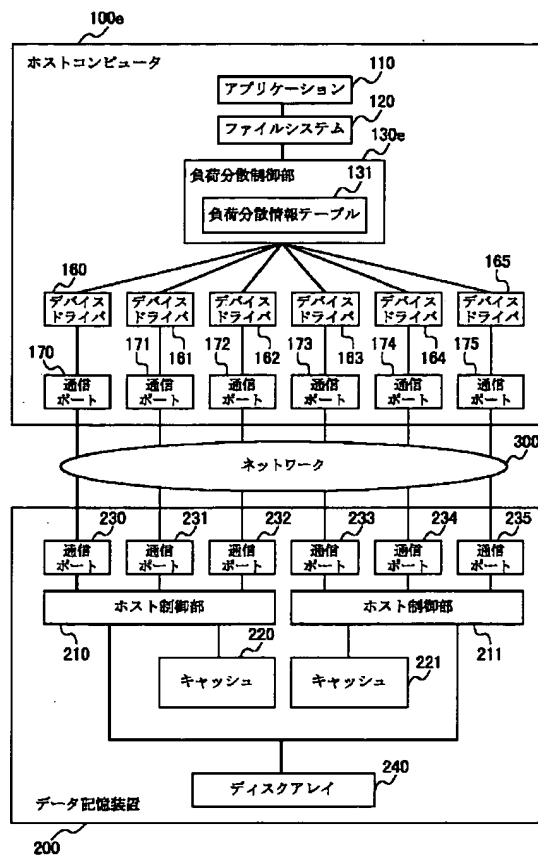
【図11】



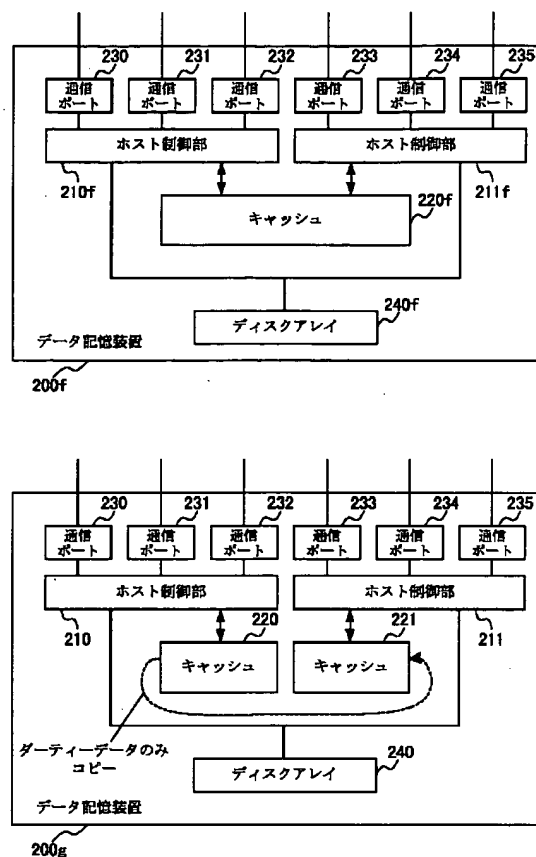
【図10】



【図12】



【図13】



フロントページの続き

(51) Int. Cl.<sup>7</sup>

G 0 6 F 12/08

識別記号

5 5 1

5 5 7

F I

G 0 6 F 12/08

7-711-1 (参考)

5 5 1 Z

5 5 7

